Iliad and Medical HouseCall: Evaluating the Impact of Common Sense Knowledge on the Diagnostic Accuracy of a Medical Expert System

Omar Bouhaddou, Ph.D., Joseph G. Lambert, M.D., G. Eric Morgan
Applied Medical Informatics
Salt Lake City, Utah

Diagnostic expert systems are gaining acceptance among physicians. Recently, a comparative study of the performance of four major commercial diagnostic programs demonstrated that the information they produce contains a certain amount of irrelevancy that the trained physician ignores. Medical HouseCall is a consumer health information expert system based on a medical expert system for physicians, Iliad. To enhance the usefulness of Medical HouseCall to health care consumers, we are interested in significantly reducing the amount of irrelevancy contained in the diagnostic differential list. Testing with over 470 'textbook' cases revealed that a large part of the irrelevancy can be eliminated by adding universal and medical 'common sense' knowledge. Using four performance measures, we compared, on a subset of cases, the differential lists from two versions of the program: the refined knowledge base (1995) and an older version (1994) 'pre-common sense'. The results suggest that the performance of a diagnostic expert system can be significantly improved with the addition of common sense knowledge.

INTRODUCTION

Medical diagnostic programs have been available commercially for the last 10 years [8]. The increased popularity of these systems within the medical community suggests that these systems perform better, are more useful, and that the attitude towards decision support system in general is more accepting. The performance of medical diagnostic systems has been evaluated in the past [4,9,10]. One recent study published in the New England Journal of Medicine suggests that the performance of diagnostic expert systems, while not consistently accurate and relevant, still showed promise [3]. The authors concluded that these programs should be used by physicians who can identify and use the relevant information and ignore the irrelevant information these systems can produce.

Can the performance of these systems be improved and their usefulness increased for the inexperienced user who may not be able to discriminate relevant from irrelevant information? Can the irrelevant information be reduced or eliminated? These are the important questions we faced as we created the first medical diagnostic expert system for the consumer. Medical HouseCall [1] is a family diagnostic software and medical encyclopedia whose expert system knowledge base stems from the years of work developing Iliad [2], a diagnostic and treatment software for physicians. Medical HouseCall offers several functions, namely: symptom analysis, drug-drug interaction, a medical record, and a medical encyclopedia. In this paper, we focus on the diagnostic module supporting the symptom analysis. The function of the diagnostic module is to help the patient express their symptoms in a precise and comprehensive manner. It takes the consumer through extensive questioning to alert them to the importance of some details regarding their illness which they can then relate to their health care provider. The list of possible causes the program generates helps relieve worries and indicates when a possible condition is serious or warrants a visit to a health care provider. But, since the consumer may not be able to identify irrelevant information, we are interested in knowledge engineering strategies that would significantly reduce the amount of irrelevancy contained in the differential list and improve its usefulness to the consumer.

After testing the diagnostic module of Medical HouseCall with more than 470 test cases and making notes of reasons why irrelevant diagnoses were included in the differential list, we empirically learned that a large part of our problem could be solved by adding general and medical 'common sense' knowledge. For instance, an erythematous lesion accompanied by scaling and pruritis and localized to the feet is not tinea capitis (ringworm of the scalp). Similarly, a female cannot have eclampsia unless she is both pregnant and has convulsions.

In this paper, we report our research efforts aimed at evaluating the impact of common sense knowledge on the performance of the diagnostic module. We entered a set of cases both in the current (1995) version of Medical HouseCall knowledge base enhanced with the addition of 'common sense' and in a year-old version of the knowledge base (1994). For each case, the two lists of diagnoses were compared in terms of 4 performance scores to evaluate the reduction in irrelevancy and any other potential improvements.

0195-4210/95/$5.00

METHODS

Background on Iliad and Medical HouseCall
Iliad is a diagnostic and treatment software for health care providers. Its current disease database covers Internal Medicine, Neurology, Pediatrics, Ob/Gyn, Psychiatry, Dermatology, Peripheral Vascular Diseases, Urology, and Sports Medicine. Iliad uses Bayesian logic combined with If-Then rules to incorporate expert judgment. The program includes an extensive term proprietary medical vocabulary to describe symptoms, physical examination findings, and test / procedure results. Iliad provides diagnostic and treatment consultation services for real cases, creates artificial cases to train and test medical problem-solving skills, and offers access to an electronic library of disease information, literature abstracts, and medical photographs. Studies have shown that Iliad is an effective teaching tool [6], a diagnostic error detection tool [4], and a surgery preauthorization tool [5].

Medical HouseCall is the consumer counter-part of Iliad. As a family medical software, Medical HouseCall allows a user at home to describe their symptoms and receive an analysis showing the list of possible causes ranked by order of likelihood. Also, a consumer can check for drug interactions among concurrent medications. In addition to these interactive features, Medical HouseCall supports medical record keeping and offers access to a large medical encyclopedia of text and photographs with hypertexting capabilities. Medical HouseCall has been reviewed in consumer magazines and presented at MedInfo'95 [1].

The Diagnostic Module
The diagnostic module (i.e., inference engine and knowledge representation models) is essentially the same in both Iliad and Medical HouseCall. However, several adjustments were made to adapt this module designed for physicians, to consumers. Some of the important adjustments are: 1) Iliad uses symptoms, physical exam findings, and lab, but only symptoms and historical information can be entered in Medical HouseCall. However, to enhance the performance Medical HouseCall, we added to the history data a large number of physical examination findings recognized as patient-observed physical findings (e.g., red throat, red eye, swollen lymph nodes); 2) all history information was translated into consumer language. For example, dyspnea was replaced with shortness of breath, polydipsia with increased thirst, acute myocardial infarction with heart attacks. However, term-to-term translation was not always sufficient and many items had to be broken into their component concepts to facilitate understanding. For instance, we decomposed the question 'smoking history in number of pack/years' into two questions: smoking history for 'number of years' and smoking history for 'number of packs per day' and decomposed joint pain into knee pain, wrist pain, etc.; 3) diagnostic probabilities were adjusted to never exceed 70%. This value has been estimated to represent the average contribution of historical information toward making a diagnosis [7]; and 4) diagnoses were combined in broader categories when the differentiation is only possible based on physical examination and lab values. For instance, Medical HouseCall has one meningitis, one pneumonia, one colitis, etc., as opposed to differentiating viral meningitis, meningococcal meningitis, etc.

In this study, we are interested to evaluate the impact of a different set of changes made to the Medical HouseCall knowledge base between the first release in April 1994 and the latest release dated March 1995. These changes represent for the most part general medical knowledge and are referred to as 'unperceived' or 'common sense' knowledge.
Common sense rule out criteria were added systematically to every diagnostic profile. If present, rule out information rules out the diagnosis. Rule outs are either coarse filters such as 'symptoms of systemic involvement' or fine filters consisting of one symptom (e.g., fever) or a modifier of a symptom (e.g., generalized rash). We introduced a datadriver flag. A data driver flag assigned to a diagnostic criteria indicates that the diagnosis should not be displayed unless that criteria is present. However, the diagnosis remains on the internal differential and will be used by the system when determining workup suggestions. For instance, Reiter syndrome appears on the list of possible causes only in the presence of eye, urinary tract, and joint abnormalities. We made extensive use of the risk factor flag. A risk factor flag assigned to a diagnostic criteria indicates that the diagnosis should not be displayed when that criteria alone is present. Thus, all personal and family history were assigned a risk factor flag and are taken into account only if some other evidence for the diagnosis is entered. The computation of the 'OR' construct was redefined. The 'OR' statement groups multiple diagnostic criteria that are inter-dependent and only allows the criteria with relatively more predictive value to be used. The 'OR' structure allows a better compliance with the independence assumption imposed by the Bayesian model. Thus, the location of a rash can be multiple, but the occurrence of the rash in multiple locations should not translate in added confidence in the diagnosis. The 'OR' statement is used to group the different locations with their respective frequencies. In the past, if one location was more specific than the others and the less common location was observed in a particular case, the finding penalized the diagnosis

because the absence of the common location weighted more than the presence of the rare location. A new way of processing the 'OR' changed this by requiring the program to always use the positive information in an 'OR' group and to ignore the negative information even if it had a higher predictive value.

These refinements to the diagnostic module required less expert knowledge than common medical knowledge. They constitute unformulated general knowledge that a human expert unconsciously uses to eliminate obvious irrelevancies or refine competitive hypotheses on the differential diagnosis list.

Description of the Test Cases

Over 470 test cases were developed to test the Medical HouseCall knowledge base. These cases are 'classic' or 'textbook' presentations of diseases, the symptoms being extracted from standard current medical texts. As pointed out in [11], these cases may be infrequently encountered. Both simple and complex disease were represented and the symptom lists for these diseases ranged from 1 to 26 symptoms. Simple, straight forward diagnoses such as meningomyelocele and hypospadias consisted of few symptoms while complex diseases such multiple sclerosis, dermatomyositis, and tuberculous sclerosis had a larger numbers of symptoms. A 'random' subset of 48 cases was used in this study. This subset represent a cross section of medical specialties: Dermatology-10%; ENT-4%; Genetics-8%; Internal medicine-40%; Neurology-2%; Ob/Gyn-8%; Ophthalmology-2%; Pediatrics-14%; Surgery-4%; and Urology-4%.

Age: 26
Sex: Female
headache
over most of the head (generalized)
severe pain
occurs repeatedly or frequently (recurrent)
visual loss or blindness
partial
temporary (transient)
began suddenly
double vision (diplopia)
bladder control - unable to release urine from the bladder (retention)
bladder control - unable to keep from leaking urine (urinary incontinence)
weakness - decreased muscle strength
hips, legs, or feet
right leg only
numbness or tingling (paresthesia)
legs or feet
right leg
hands and feet (distal extremities)
movement - loss of movement (paralysis)
only one side of the body (hemiparesis)
right leg
[No] hips, legs, feet (lower extremities)
[No] developed in a bottom-to-top pattern (ascending paralysis)
[No] developed in a top-to-bottom pattern (descending paralysis)
[No] occurs on both sides equally (symmetrical)
movement - lack of coordination in muscular movements (incoordination)
language - impaired speech or language slurred speech

Figure 1: Example of a 'classical' case for multiple sclerosis symptoms.

Analyses of Cases

To study the effect of 'common sense' on the performance of the diagnostic module, we have developed and adopted from others [3] the following measures. For each test case a differential diagnosis list was generated using the April 1994 and March 1995 versions of Iliad programs and knowledge bases. Each disease on the lists is associated with a post-test probability or posteriori value (in this study, the posteriori was not normalized to 70%). These lists contained 20 or less entries as the diagnostic module limits the display to 20 or to those diagnoses that are above 1% likelihood whichever is smaller. The two lists of diagnoses were reviewed without blinding by a physician (JGL) who assigned a 1 to each diagnosis he judged as an appropriate hypothesis (relevant).

The score for correct diagnosis is calculated on the number of appropriate diagnoses divided by the total number of diagnoses on the differential. The score for irrelevant diagnoses is based on the sum over all irrelevant diagnoses on the differential of {(number of diagnoses on the list + 1 - rank) * posteriori}. Thus, the higher an irrelevant diagnosis is on the differential list, the higher is its likelihhood, and the higher is the irrelevancy score. To normalize the value obtained, we divided the score by the maximum value calculated as {(number of diagnoses on the list) + (number of diagnoses on the list-1) + ... + 1} which is equal to {number of diagnoses on the list * (number of diagnoses on the list + 1) / 2}. This final score is normalized and varies between 0% and 100%. The score for rank is the average rank of the correct diagnosis as it appears on the differential list. The score for posteriori is the average posteriori of the correct diagnosis as it appears on the differential list.

Statistical Analysis

For each version of the program, we calculated means and 95 percent confidence intervals. The difference between the means from two versions of the program were tested for statistical significance with a t test for paired observations.

RESULTS

Table 1 shows the scores obtained by the 1994 and 1995 versions of the program on the subset of 48 test cases.

Table 1: Performance scores of Iliad/Medical HouseCall in 1994 and 1995.

Scores	1994	1995	t statistic (P-value)
total number of test cases	48	48	
average number of diagnoses on the differential	15.6	7.3	12.115 (.0001)
average irrelevancy score	.21	.05	6.027 (.0001)
average correct diagnosis score	.19	.43	9.442 (.0001)
average rank of correct diagnosis	7.3	1.2	5.069 (.0001)
average probability of correct diagnosis	.40	.83	7.577 (.0001)

The same 48 cases were used to test the 1994 and 1995 versions of the program. Since the test cases were created using the 1995 version, a number of dictionary terms were undefined in the 1994 version (an average of 6% per case). It was not always a case of our not representing those symptoms in the 1994 version, rather, that we had changed their coded location in the dictionary. To correct for this bias, the corresponding symptoms, as represented in the 1994 knowledge base dictionary, were added to the 1994 case to standardize the comparison. In instances where the symptoms were missing from the 1994 knowledge base, the cases containing these symptoms were discarded from the study.

The revisions made in 1995 shortened the length of the differential from an average of 15.6 to 7.3 entries. Because the shortening of the differential list was accompanied by an increase in the number of correct diagnoses and a decrease in the amount of irrelevancy, the resultant list of possible causes is more reliable and easier to review. The irrelevancy score decreased from 21% to 5% between 1994 and 1995, in other words, the relevance of the differential diagnosis list increased. In general, the irrelevancy score is underestimated because of the maximum value used to normalize it. Indeed, we estimated that the maximum score is reached when all diagnoses on the differential are incorrect (irrelevant) and their posterior probability is 1. The proportion of correct diagnoses has increased from 19% to 43% respectively. This also demonstrates increased performance of the diagnostic module. The rank score of the correct diagnosis decreased from 7.3 to 1.2 and the posteriori score increased from .40 to .83. In other words, the correct diagnosis was on average listed in the middle of the differential list in 1994 whereas in 1995, it was listed in the top 2. Also, on average, the correct diagnosis was reached with more confidence in 1995 (83%) than in 1994 (40%). Overall, all comparisons were statistically significant at the 0.02% level .

DISCUSSION

As Eta Berner said about the evaluation of computer-based diagnostic systems, two major issues need to be addressed: accuracy and usefulness [3]. This study addresses only the first, although, clearly the impact carries to the second, since as accuracy increases usefulness increases. However, because of the serious biases existing in this study and discussed below, we are only looking at the relative performance between two versions of the same program and the benefits of adding 'common sense' knowledge to a diagnostic expert system.

The subset of 'classical' cases used in this study are too few and too artificial to constitute a definitive measure of the diagnostic accuracy of any system. These cases have been built by the knowledge engineers to test and debug the 1995 version of the diagnostic module. Therefore, we have no guaranties that performance of the system on these cases simulates performance of real users in the field. The assignment of the 'relevant' status is also subjective since it was done by the author of the cases (JGL) and without blinding. However, this bias is equally represented in both versions and therefore does not advantage one version in particular. The difference in the number of entries in each diagnosis list directly influences the calculation of the proportion of correct diagnoses and the irrelevancy score. However, if the lists of entries in the 1995 differentials were shorter it is because less diagnoses reached a posteriori higher than the 1% cutoff point. If we had required each list to have 20 entries, we would have added to the 1995 lists entries with 1% or less likelihood. This would not have impacted too significantly the numerator of

the irrelevancy score but would have increased significantly the maximum value denominator. As a result, we would have a decrease in the irrelevancy score and an even greater statistical difference between the two versions than reported here.

Although the proportions of correct diagnoses shown in Table 1 are not impressive in both 1994 and 1995 performances, in 1995 the correct diagnosis was included in the differential list 100% of the time and 85% of the cases included the correct diagnosis at the top of the differential list. Moreover, the other diagnoses are listed with an indication of their relative likelihood which is usually low. This suggests that the diagnostic module has been improved to recognize the correct diagnosis more often and with more confidence but that the differential still includes some 'noise', especially at the bottom of the list in hypotheses at less than 5% likelihood.

The acceptability of computer-based diagnostic systems will only be considered if system performance improves and usefulness increases. In the hands of a novice, the work of eliminating irrelevant diagnoses from the list of possible causes is even more urgent, otherwise the use of the system could lead to unnecessary worries. The study presented has important biases which limit any claim regarding the absolute effect of 'common sense' or general medical knowledge on the accuracy of the diagnostic module. Also, we have not specified how to systematically include 'common sense' in a diagnostic knowledge base. However, the study suggests there is a relative improvement in the diagnostic performance due in part to adding general medical knowledge.

Acknowledgments

This study would not have been possible without the contribution of the Knowledge Engineering Teams at Applied Medical Informatics (K. Sward, R. Hulet, C. Fan, W. Wang, S. Jensen) and at the University of Utah, Department of Medical Informatics (H. Warner, D. Sorenson, H. Yu, B. Bray). The authors also acknowledge the improvements SCAMC reviewers made to this paper. This work is supported in part by a grant from NIST/ATP program.

References

1. Bouhaddou O, Warner HR Jr. An Interactive Patient Information and Education System (Medical HouseCall) based on a Physician Expert System (Iliad). Medinfo'95, Vancouver - Canada 1995.
2. Warner HR, Haug PJ, Bouhaddou O et al. Iliad as an Expert Consultant to Teach Differential Diagnosis. 13th Symposium on Computer Applications in Medical Care, pp371-376 Washington November, 1988.
3. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. New England Journal of Medicine, June 23 1994;330(25):1792.
4. Lau Lee Min and Warner HR. Performance of a Diagnostic System (Iliad) as a Tool for Quality Assurance. 15th Symposium on Computer Applications in Medical Care, Washington November, 1991.
5. Bouhaddou O, Frucci L, et al. Implementation of Practice Guidelines in a Clinical Setting using a Computerized Knowledge Base (Iliad). 17th Symposium on Computer Applications in Medical Care, Washington November, 1993.
6. Lincoln MJ, Turner CW, Haug PJ et al. Iliad's role in the generalization of learning across a medical domain. 16th Symposium on Computer Applications in Medical Care, Washington November, 1992.
7. Peterson MC, Holbrook JH, De Von Hales et al. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. The Western Journal of Medicine. 1992 eb; 2(156):163-166.
8. Miller RA. Medical diagnostic decision support systems - past, present, and future. J Am Med Informatics Assoc. 1994;1:8-27.
9. Feldman MJ, Barnett GO. An approach to evaluating the accuracy of DXplain. Comput Methods Programs Biomed 1991;35:261-6.
10. Heckerling PS, Elstein AS, Terzian CG, Kushner MS. The effect of incomplete knowledge on the diagnosis of a computer consultant. Med Inf (Lond) 1991;16:363-70.
11. Dell S. Why are classic cases rare? New England Journal of Medicine, July 24 1980;303:228.